Gene Expression Analysis by DNA Computing

Akira Suyama[1,2]
suyama@dna.c.u-tokyo.ac.jp
Nao Nishida[2]
nishida@genta.c.u-tokyo.ac.jp
Ken-ichi Kurata[1]
kurata@genta.c.u-tokyo.ac.jp
Katsumi Omagari[2]
omagari@genta.c.u-tokyo.ac.jp

[1] Institute of Physics and
[2] Department of Life Sciences, Graduate School of Arts and Sciences,
University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Keywords: functional genomics, DNA computing, gene expression profile, DNA chip

1 Introduction

DNA chips perform the rapid, parallel detection of a target set of gene transcripts through specific hybridization of targets to DNA probes integrated on the chip surface [1]. A DNA chip can thus be considered as a special purpose DNA computer for parallel homology search. Conventional DNA chips rely solely on hybridization for transcript detection. As a result, the recognition process for genome scale gene expression profiling (GEP) requires the use of very expensive, target-dependent, high-density DNA chips. In this paper we described more sophisticated but easy-to implement DNA computing (DNAC) algorithm for GEP. This algorithm does not require the use of a target-dependent high-density DNA chip. Indeed, GEP is achieved using a universal DNA chip with a small number of DNA probes, whose sequences have been optimized for accurate and quantitative hybridization.

2 Methods and Results

The proposed DNAC algorithm for GEP consists of three processes: *ENCODE*, *AMPLIFY* and *DECODE* (Fig. 1). The *ENCODE* process constructs a set of single-stranded (ss) DNA molecules, which make up a GEP table, from the transcripts present in a tube of interest, T. First, a partially double-stranded (ds) DNA adapter molecule (A_i) is made for each target transcript species, i. A_i contains a ss region, s_i, complementary to a unique subsequence of target transcript i, and a ds region encoding a unique DNA-coded number, DCN_i. In a base-n implementation, each DCN_i consists of a pair of sequences, $D1_i$ and $D2_i$, which is flanked by the common sequences SD and ED. Extraction of the DCN which corresponds to each transcript in T is facilitated by transcript-mediated ligation with a biotin-labeled sequence, e_i. This operation is identical to the *append* operation, which has been used to solve a 3-SAT instance using a DNA-implemented dynamic programming algorithm [4]. All adapter molecules captured on streptavidin (SA) magnetic beads are then melted into single strands to obtain a set of ssDNA molecules representing a GEP table.

Those sequences constituting DCN_i are chosen from a set of orthonormal sequences, which have uniform T_m, and no mishybridization or folding potential. A set of over 200 orthonormal sequences of length 25 nt has been designed using a greedy algorithm [4]. This set is sufficient to uniquely encode over 10^4 distinct transcripts into 2-digit, base 100 DCNs.

The *AMPLIFY* process uniformly amplifies the GEP table. ssDNA species in tube T^* which encode for a DCN (each of the sequence $\overline{SD}$-$\overline{D1_i}$-$\overline{D2_i}$-$\overline{ED}$) are subjected to parallel amplification by PCR. The uniformity of amplification is facilitated by the use of the common primer sequences SD and

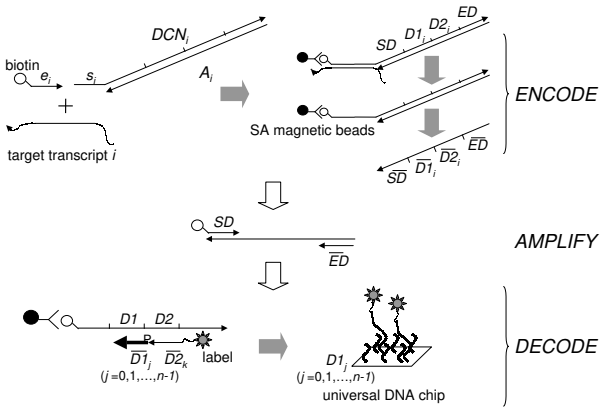

Figure 1: DNA computing algorithm for gene expression profiling.

$\overline{ED}$ and the orthonormality of sequences $D1_i$ and $D2_i$. Such uniformity is difficult to achieve when amplifying mRNA transcripts directly.

The *DECODE* process serves to display the contents of the amplified GEP table. This procedure is carried out in two steps. First, the amplified tube T^* is divided into n identical tubes, $T_0^*, \cdots, T_{n-1}^*$. A fluorescent-labeled ssDNA species complimentary to sequence $D2_k$ is then added to each tube T_k^*, for $k = 0$ to $n-1$, and allowed to hybridize. ssDNA molecules encoding for the complement of each possible first digit are also added to every tube and allowed to hybridize. Each tube is then subjected to ligase. Every ligated ssDNA molecule from tube T_k^* must encode for $D2_k$. The set of first digits present for each second digit, $D2_k$, may then be determined by exposing the ligated ssDNA molecules in tube T_k^* to a universal DNA chip, to which are attached DNA probe sequences representing all first digits ($D1_j, j = 0$ to $n-1$). A GEP is finally obtained by combining the hybridization profiles of all n tubes. A DNA capillary array is indeed suitable for parallel hybridization of n tubes in the *DECODE* process [3].

A simple experiment was performed to investigate the feasibility of this DNAC algorithm for GEP. Two ssDNA oligonucleotides of length 30 nt were designed to represent unique sequences of the mRNA transcripts, IGTP and LRG-47, from the mouse genome. The design was performed using the efficient algorithm for finding out a set of unique sequences of target transcripts [2]. It was demonstrated that each DNAC process illustrated in Fig. 1 was executed in a highly specific and quantitative manner.

Acknowledgements

The authors thank Dr. John A. Rose for critically reviewing the manuscript. This work was supported in part by a grant from the Japan Society for the Promotion of Science "Research for the Future" Program (JSPS-RFTF96100101) and NEDO (New Energy and Industrial Technology Development Organization) (SW, YM and YE).

References

[1] Lander, E.S., Array of hope, *Nature Genetics*, 21:3–4, 1999.

[2] Kurata, K. and Suyama, A., Probe design for DNA chips, *Genome Informatics*, 10:225–226, 1999.

[3] Suyama, A., DNA chips – Integrated chemical circuits fro DNA diagnosis and DNA computers, *Proc. 3rd International Micromachine Symp.*, 7–12, 1997.

[4] Yoshida, H. and Suyama, A., Solution to 3-SAT by breadth first search, *Proc. 4th International Meeting on DNA Based Computers*, 9–20, 1999.